

# Tuning target selection algorithms to improve galaxy redshift estimates

Ben Hoyle<sup>1,2</sup>, Kerstin Paech<sup>1,2</sup>, Markus Michael Rau<sup>1,3</sup>,  
Stella Seitz<sup>1,3</sup>, Jochen Weller<sup>1,2,3</sup>

<sup>1</sup> *Universitaets-Sternwarte, Fakultae fuer Physik, Ludwig-Maximilians Universitaet Muenchen, Scheinerstr. 1, D-81679 Muenchen, Germany*

<sup>2</sup> *Excellence Cluster Universe, Boltzmannstr. 2, D-85748 Garching, Germany*

<sup>3</sup> *Max Planck Institute fuer Extraterrestrial Physics, Giessenbachstr. 1, D-85748 Garching, Germany*

E-mail: hoyleb@usm.uni-muenchen.de

Accepted —. Received —; in original form —.

## ABSTRACT

We showcase machine learning (ML) inspired target selection algorithms to determine which of all potential targets should be selected first for spectroscopic follow up. Efficient target selection can improve the ML redshift uncertainties as calculated on an independent sample, while requiring less targets to be observed. We compare 7 different ML targeting algorithms with the Sloan Digital Sky Survey (SDSS) target order, and with a random targeting algorithm. The ML inspired algorithms are constructed iteratively by estimating which of the remaining target galaxies will be most difficult for the machine learning methods to accurately estimate redshifts using the previously observed data. This is performed by predicting the expected redshift error and redshift offset (or bias) of all of the remaining target galaxies. We find that the predicted values of bias and error are accurate to better than 10-30% of the true values, even with only limited training sample sizes. We construct a hypothetical follow-up survey and find that some of the ML targeting algorithms are able to obtain the same redshift predictive power with 2-3 times less observing time, as compared to that of the SDSS, or random, target selection algorithms. The reduction in the required follow up resources could allow for a change to the follow-up strategy, for example by obtaining deeper spectroscopy, which could improve ML redshift estimates for deeper test data.

**Key words:** galaxies: distances and redshifts, catalogues, surveys.

## 1 INTRODUCTION

In order to maximise the cosmological information from large scale structure surveys, samples of galaxies must be identified and their positions on the sky, photometric properties, and redshifts measured. Measuring accurate spectroscopic redshifts is resource intensive and is typically only performed for a small sub sample of all galaxies. For this sub sample of galaxies one may learn a mapping between the measured photometric properties (or ‘features’), and the spectroscopic redshift. The learnt mapping can then be applied to all photometrically identified galaxies to estimate photometric redshifts.

This paper aims to address the problem of identifying which galaxies, from an available target list, should be targeted first for spectroscopic follow up, in order to reduce the uncertainty on the machine learning redshift estimates of a final test sample. To examine this problem, and to test our

methods, we construct sets of simulated (or hypothetical) observing runs using existing data.

Depending on the science case, one may wish to target particular types of galaxies for spectroscopic follow-up. The authors Jouvel et al. (2014) use artificial Neural Networks in a classification analysis to select potential targets using photometric data, that are consistent with being emission line galaxies.

Within the science case of attempting to improve the photometric redshifts of sub samples of galaxies, (Carrasco Kind & Brunner 2013) suggest a method to use random forests to estimate photometric redshift pdfs. The authors identify volumes in color-magnitude space which produce poor redshift estimates, as defined by the compactness of the photometric redshift probability distribution function (hereafter pdf). They suggest that targeting galaxies in these color cells can lead to improvement in the photometric redshift of other galaxies within the cell. The authors (Masters

et al. 2015) use a similar technique by compressing a high dimensional color space into complex two dimensional cells using Self Organised Maps (Kohonen 1997). The authors determine which regions of the projected feature space have cells which are populated by a test sample galaxy, but are not currently populated by a spectroscopic training example. Underpopulated cells can then be targeted for follow up, however determining the density of spectra in a cell is difficult because of the non trivial cell volume. The method presented in this paper differs from these approaches. We attempt to predict which galaxies should be targeted next, and in which order, from a list of all possible targets, such that the redshift metrics on a final test sample are iteratively improved. This differs from the above approaches by not requiring that spectra are taken in certain feature cells, if they are not required to improve redshift estimates, or that some cells could still be targeted further if this helps the remaining sample to have improved redshift estimates. We also investigate different targeting algorithms and compare their performance using standard redshift performance metrics.

Machine learning methods and techniques have long been applied to photometric redshifts analysis since Tagliaferri et al. (2003) used artificial Neural Networks. Since this time a plethora of machine learning architectures, including the tree based methods used in this work, have been applied to the problem of point prediction redshift estimation or to estimate the redshift pdf (see e.g., Collister & Lahav 2004; Csabai et al. 2007; Carliles et al. 2008; Gerdes et al. 2010; Bonnett 2015; Carrasco Kind & Brunner 2013; Hogan et al. 2015; Rau et al. 2015). Further concepts have also recently been ported from machine learning to astronomy, such as feature importance (Hoyle et al. 2015), feature extraction (Polsterer et al. 2014), data augmentation (Vanzella et al. 2004; Hoyle et al. 2015) and deep machine learning (Hála 2014; Dieleman et al. 2015; Hoyle 2015). Machine learning architectures have also had success in other fields of astronomy such as galaxy morphology identification, and star & quasar separation (see for example Lahav 1997; Yeche et al. 2009). Other data driven, albeit non machine learning, redshift estimates techniques exist and are gaining popularity (e.g., clustering techniques Menard et al. 2013).

The original target selection algorithm for the Sloan Digital Sky Survey (hereafter SDSS, York & SDSS Collaboration 2000) was proposed to fulfill a broad range of criteria, including; the uniformity of the sample, the insensitivity to systematic effects such as seeing, the requirement that it should be based on physically meaningful parameters which correlated with galaxy properties, and that it should be as simple as possible to allow for the construction of mock catalogues (Strauss et al. 2002). Afterwards other targeting algorithms were suggested to construct samples of Luminous Red Galaxies (Eisenstein et al. 2001) and Quasars (Richards et al. 2002). This paper uses SDSS data to explore other targeting algorithms, which are tuned to reduce the redshift uncertainty of an independent test sample of representative galaxies.

This paper is organized as follows: In §2 we describe the data sample and the machine learning methods employed, the design of the experiments and their analysis in §3, and results in §4. We summarise and conclude in §5.

## 2 DATA AND MACHINE LEARNING METHODS

In this study we use observational data drawn from the SDSS III Data Release 12 (Alam et al. 2015), and divide the sample into two sub groups by the date when the spectra were taken. These sub samples correspond to approximately SDSS I&II and SDSS III. We analyse each sub sample separately during independent sets of analyses.

### 2.1 Observational dataset

The SDSS I-III uses a 2.5 meter telescope at Apache Point Observatory in New Mexico and has CCD, wide field photometry, in 5 bands ( $u, g, r, i, z$  Gunn et al. 2006; Smith et al. 2002), and an expansive spectroscopic follow up program (Eisenstein 2011) covering  $\pi$  steradians of the northern and equatorial sky. The SDSS collaboration has obtained more than 3 million galaxy spectra using dual fiber-fed spectrographs. An automated photometric pipeline performs object classification to a magnitude of  $r \approx 22$  and measures photometric properties of more than 100 million galaxies. The complete data sample, and many derived catalogs such as galaxy photometric properties, are publicly available through the **CasJobs** server (Li & Thakar 2008)<sup>1</sup>.

The SDSS dataset is well suited for the analyses presented in this paper due to the large number of photometrically selected galaxies with spectroscopic redshifts to use as training and test samples and the documented date corresponding to when the spectra were taken. These spectral dates are used as one of the comparison target selection algorithms, described further in §3.

We select all objects from **CasJobs** with both spectroscopic redshifts and photometric properties. In detail we run the **MySQL** query shown in the appendix resulting in 3,751,496 objects. We next select all of the 2,183,897 objects which are classified by the photometric pipeline **PHOTPTYPE** as being a galaxy, and have measured errors on the model magnitudes in all  $g, r, i$  bands to be less than 0.2. We next select all of the 2,158,880 objects with a spectroscopic redshift error below 0.001, and with spectroscopic redshift to be above 0, and to have no spectral warning flags set. Finally we remove duplicate photometric **objid** which reduces the sample to 1,981,397. This selects a clean sample of galaxies, free of stellar contamination, which is suitable for the analysis in this work. One may also choose to select combinations of stars, galaxies, and other artefacts if the science goal were to improve star/galaxy separation. We note that stellar contamination, or contamination by AGN is a very real problem which plagues target selection routines and photometric redshift estimates, especially at fainter magnitudes. However while non galaxy sources may contaminate the galaxy sample in reality, they are important to observe, to construct a representative samples in order to accurately perform star, AGN and galaxy separation, in combination with accurately estimating redshifts.

From this base set we construct two samples, based on the MJD time stamp that the galaxy spectra was taken. We select 869,479 objects to approximately construct the

<sup>1</sup> [skyserver.sdss3.org/CasJobs](http://skyserver.sdss3.org/CasJobs)

SDSS I&II sample by selecting spectra with MJD values to be less than 5.47e4, and 1,111,918 spectra with MJD values greater than 5.47e4 to be the SDSS III sample. We randomly extract 100,000 galaxies from both samples to form the list of potential training set target galaxies, and use the remaining galaxies as the final test sets to estimate redshift prediction ability in each set of analyses. We have also explored using 80% of the galaxy sample as the potential target galaxies and find very similar results. However using these larger samples leads to a dramatic increase in computation time due to the number of machine learning systems we construct for each of the hypothetical observing runs, as described in §2.2 and §3.

In this work we have concentrated on the following six typical features for redshift estimation; the  $r$  band magnitude, the following colours:  $r - g$ ,  $g - u$ ,  $i - r$ ,  $z - r$ , and the Petrosian radius measured in the  $r$  band. Previous work has shown that there are many other readily obtained photometric features which also have strong predictive power when estimating redshifts (Hoyle et al. 2015), however defining an optimal redshift estimation technique for the data sample in this paper is not the main focus of this work.

## 2.2 Tree based methods

We use the scikit-learn implementation of decision trees for regression (Breiman et al. 1984) as the machine learning architecture to predict galaxy redshifts, redshift errors, and redshift biases. The decision tree algorithm recursively partitions the input feature dimensions into an increasing number of bins. The bin boundaries are chosen to minimize the scatter of the output feature for all of the object which sit in each bin. In this work the output feature can be the spectroscopic redshift, the photometric redshift bias, or the photometric redshift error (see §3).

The power of tree based methods is enhanced by combining many trees. One technique to do this is called Random Forests (Breiman 2001) which constructs  $N$  trees simultaneously by drawing random selections of both feature space and random (with replacement) samples of training data  $D_{tr}$ , with which to train each tree. Each tree is grown by recursively partitioning the selected property (or ‘feature’) with the choice of partition centre being performed using a ‘greedy’ strategy, in order to reduce the variance of the output feature between the data in each partition. The data which sit on each final leaf node form the prediction value for new data  $D$ , which is queried down the tree. Each tree  $T$ , can be viewed as learning a model of the data, but is prone to either over fitting, or not modelling the complexity in the data well enough. However, the combined prediction from many trees produces a forest prediction  $P$  which is a model with strong predictive power, and is not prone to over fitting. The random forest prediction is obtained for new data  $D$  following

$$P(D) = \frac{1}{N} \sum_{i=0}^N T_i(D). \quad (1)$$

In a regression analysis, such as that used to predict redshift estimates, the final output value is therefore the average value from each tree in the forest. We also measure the

standard deviation of predictions from all of the trees for each galaxy  $g$ , and refer to this quantity as the photometric error, defined as  $\delta_{z_{ML}} = \sigma(T(D_g))$ .

For more details about constructing random forests we refer the reader to Hastie et al. (2001)<sup>2</sup>. The hyper-parameters of the random forest include the final number of data on each leaf node  $n_L$ , the number of trees in the forest  $N$ , and the maximum number of input features that are separately selected during the training of each tree in the forest,  $n_F$ .

In all of the work which follows, and when referring to machine learning systems, we will imply the use of random forests. For each random forest system, we actually construct eight random forests, each with a random choice of hyper-parameter values, and we select the best fitting random forest for the task at hand by using a cross validation hold-out sample. We do not perform a grid search of hyper-parameter space due to the large number of systems which are trained. We normally find that after training 30 machines with randomly selected hyper-parameters, the best trained machine produces metric values which are typically very close to that of a comparable system, but with more than 100 random hyper-parameter selections. Our choice of 8 random samples of hyper-parameter ensures that we still explore the space well, but not exhaustively, given the number of systems that need to be train. We also note that the hyper-parameters of the best fitting model is always non trivial to predict a priori. We define the best random forest to be the one with the smallest value of the harmonic mean of the 68% dispersion (denoted  $\sigma_{68}$ ) and the 95% dispersion ( $\sigma_{95}$ ) of the measured quantity. The harmonic mean is defined by  $\sigma_{68} \times \sigma_{95} / (\sigma_{68} + \sigma_{95})$ . We choose to randomly explore the following forest hyper-parameters during this process; the number of trees in the forest from  $30 < N < 250$ , the minimum number  $n_L$  of data on each leaf node of the tree to be between  $1 < n_L < 100$ , and the maximum number of features  $n_F \leq 6$  which may be chosen during the random feature selection.

## 3 METHOD AND ANALYSIS

In this section we define and motivate the different target selection algorithms used in this work. In §3.1 we build a system to predict which targets will be the most difficult to accurately estimate a redshift, and quantify this prediction process in §3.1.2. We describe how the different target selection algorithms use these predictions to identify samples for follow up in §3.2. We finally describe how the simulated observing runs are performed in §3.3.

### 3.1 How to identify poor targets

To identify which targets will be the most difficult to accurately estimate a redshift we use machine learning systems in two distinct stages, which we outline here, and describe in more detail in §3.1.1.

This method is somewhat different from a standard machine learning approach, in which one attempts to estimate

<sup>2</sup> [statweb.stanford.edu/~tibs/ElemStatLearn](http://statweb.stanford.edu/~tibs/ElemStatLearn)

the photometric redshift  $z_{ML}$ , of a galaxy, and to estimate the photometric redshift error  $\delta_{z_{ML}}$ . Here we rather attempt to predict how large the photometric redshift error  $\delta_{z_{ML}}$  on each galaxy will be, and also to predict the offset, or bias, between the redshift point prediction and the true spectroscopic redshift, defined as  $b_{z_{ML}} = |z_{spec} - z_{ML}|$ .

To achieve these aims, random forests are first used in a standard redshift estimation process using 85% of the training sample, which iteratively grows after each simulated observing run. The remaining 15% of the training data is then passed through this first stage and photometric redshifts are estimated. We then calculate the true redshift offset, and redshift error using the 15% training sample. We train the next machine learning system to predict these values of photometric redshift errors and offsets of this 15% training sample using only the photometric input features.

All of the galaxies in the potential target list are passed through this second system to predict and identify which targets will have large redshift errors, and large offsets. Many of these poorly performing targets are selected by the targeting algorithms and then ‘observed’ to obtain a spectroscopic redshift. A schematic diagram of this process is shown in Fig. 1.

Finally a third machine learning system is trained for redshift prediction using all of the training galaxies, including the newly observed targets. The photometric properties of the test sample are passed into this third system and redshifts estimates are obtained. We use the test sample redshift predictions to measure performance metrics, see §3.4 for details.

We do not attempt to predict the full shape of the redshift probability distribution function for each target because it would induce unnecessary uncertainty. By concentrating on the descriptive point estimates  $b_{z_{ML}}, \delta_{z_{ML}}$ , the signal to noise of the predictions are increased. Calculating these point predictions is also less computationally intensive, and requires less storage space, while also removing the need to select an appropriate band width smoothing scale required for a probability distribution function (see, e.g. Rau et al. 2015).

### 3.1.1 Predicting poorly estimated targets

Each set of predictions is remade before each simulated observing run and draws from spectra which have been observed up to, and including, the most recent run. Two training samples of size 85% and 15% are randomly constructed from this base sample. The 85%/15% sample split choice is rather arbitrary, but does ensure that most of the data is used to estimate accurate redshifts. The 15% sample is still of a reasonable size and quickly grows to be more than a few thousand objects, after just over 10% of the simulated follow up program. The first 85% of the training data is used to train a machine to estimate a completely standard photometric redshift  $z_{ML}$ , and photometric redshift error  $\delta_{z_{ML}}$  using the input photometric features  $\Theta$ , and the spectroscopic redshift  $z_{spec}$ , as shown in the upper rectangular box in the schematic diagram in Fig. 1. The 15% held out sample is passed through the learnt redshift machine to predict both a photometric redshift  $z_{ML}$ , and error  $\delta_{z_{ML}}$ . The redshift offset  $b_{z_{ML}} = |z_{spec} - z_{ML}|$  of this 15% sample is calculated exactly, by using the spectroscopic redshifts. A second

round of machine learning systems are next trained to learn the mapping between the input photometric features  $\Theta$ , and both  $b_{z_{ML}}$  and  $\delta_{z_{ML}}$  separately. For succinctness, this has been shown by the single lower rectangular box in Fig. 1.

We subsequently pass the photometric features  $\Theta$  of all potential target candidates, which have not yet been observed in our experiment, through the second machine learning system to estimate the redshift offset  $b_{z_{ML}}^{est}$ , and the redshift error  $\delta_{z_{ML}}^{est}$ , see the bottom of Fig. 1. We will next use these predictions in §3.2 to select the poorest performing potential targets using a variety of metrics. But first, we would like to present the following short interlude. Because this is a controlled experiment, and we know the true redshift of all of the targets, we can examine how well we are able to predict the two quantities: redshift offset  $b_{z_{ML}}^{est}$ , and the redshift error  $\delta_{z_{ML}}^{est}$ , of the potential target sample.

### 3.1.2 Quantifying the prediction performance

The comparison between the predicted values of offset and bias, and the true values, is performed by first passing the target’s photometric quantities  $\Theta$  through the first system to measure the photometric redshift  $z_{ML}$  and the ‘true’ redshift error  $\delta_{z_{ML}}$ . The ‘true’ offset is then constructed using  $b_{z_{ML}} = |z_{spec} - z_{ML}|$ . The photometric quantities  $\Theta$  of the targets are next passed through the second system to estimate the offset  $b_{z_{ML}}^{est}$ , and error  $\delta_{z_{ML}}^{est}$ .

Residual vectors are then constructed by subtracting the estimated value from the ‘true’ value, for all targets, e.g.,

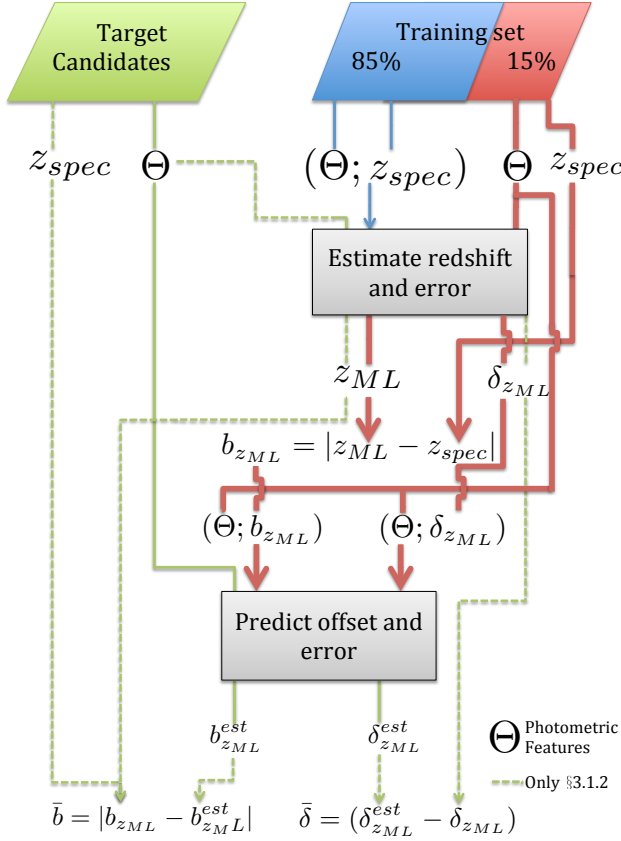
$$\bar{b} = |b_{z_{ML}}^{est} - b_{z_{ML}}|, \quad \bar{\delta} = (\delta_{z_{ML}}^{est} - \delta_{z_{ML}}). \quad (2)$$

These steps are shown on the left hand side, and by the green lines emanating from the target candidates data set, on the schematic diagram in Fig. 1. We reiterate that the targets are not used during the training of either of the two machine learning stages.

To quantify the predictive power of this process on the SDSS I&II data we measure the metrics  $\sigma_{68}, \sigma_{95}$ , on the residual vectors  $\bar{b}$  and  $\bar{\delta}$ , and also the median<sup>3</sup> of the distribution of  $\bar{\delta}$ , denoted by  $\mu$ , and present them in Fig. 2. The x-axis of Fig. 2 describes how many sets of simulated observing runs have been completed in relation to the lifetime of the survey. The data points correspond to the mean of the measured quantities and the hashed contours correspond to the 68% spread of each of the metric values as measured by the 7 different machine learning targeting algorithms, which are described in §3.2.

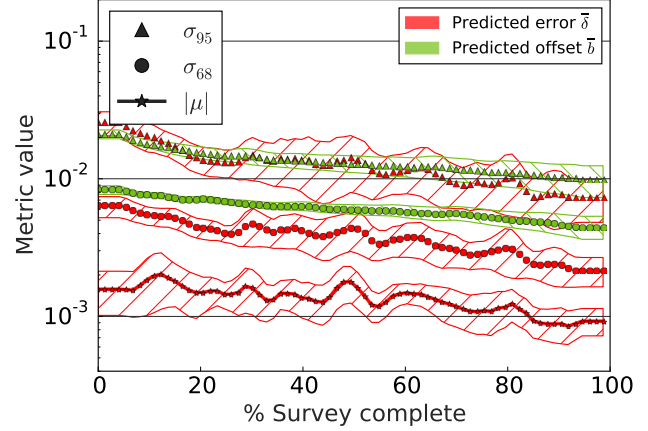
Fig. 2 shows that after each simulated observing run the ability of the systems to estimate the redshift offset  $b_{z_{ML}}^{est}$ , and the redshift error  $\delta_{z_{ML}}^{est}$ , incrementally improves. The lighter green symbols and hashed regions in Fig. 2 show that random forests are able to predict the value of the measured redshift offset to within  $9 \times 10^{-3}$  in 68% of cases, as shown by  $\sigma_{68}$  in the figure, and requires less than 13k galaxies (or 10% of the total number of simulated observing runs) to achieve this accuracy. Furthermore the 95% spread of

<sup>3</sup> Note that the median value of the predicted offsets  $\bar{b}$ , corresponds to the Median Absolute Deviation, or MAD, because of the absolute value in the definition.



**Figure 1.** A schematic diagram illustrating the prediction of the photometric redshift offset  $b_{z_{ML}}^{est}$ , and the photometric redshift error  $\delta_{z_{ML}}^{est}$ , constructed using the training data. The photometric properties of the target candidates are passed through the systems to estimate offsets and errors. Only in §3.1.2 are these estimates compared with the true values using  $\bar{b}$  and  $\bar{\delta}$ , for this controlled experiment. We refer the reader to §3.1 for a fuller description.

the predicted photometric offset as shown by  $\sigma_{95}$ , is within 0.04. If we compare these values with typical photometric redshift predictions which have an accuracy of between 0.02 and 0.05 (for the SDSS sample), we find an accuracy within about 30% of the true value. The darker red symbols and hashed regions present the prediction of the redshift error. We find that the median value of the redshift error, drops below  $2 \times 10^{-3}$  with even limited training data, and the 68% spread of predictions (as labelled by  $\sigma_{68}$ ), is below  $5 \times 10^{-3}$  within less than 15% of total observing runs, or 20k galaxies. This is very encouraging and suggests that machine learning is able to correctly estimate the size of the photometric redshift error of a representative, but unseen galaxy, to within  $(2 \pm 5) \times 10^{-3}$ , which is over an order of magnitude smaller than the true redshift error value. We note that the prediction ability depends on the choice of targeting selection algorithm because each one will follow up different galaxies during different observing runs, thus compiling a different training sample. The spread of the hashed regions show what effect this has on the measured metrics between the different targeting algorithms. We note that the SDSS III predictions are qualitatively very similar to those of SDSS I&II as shown in Fig. 2, but are approximately a factor of two times higher in each metric value. The flatness of all



**Figure 2.** The ability of the machine learning systems to predict the photometric redshift offset  $b_{z_{ML}}^{est}$ , and the photometric redshift error  $\delta_{z_{ML}}^{est}$ , of the SDSS I&II galaxies which are not used during training. The metrics (see legend) are measured on the residual vectors  $\bar{b}$  and  $\bar{\delta}$ , which are defined in the text. The x-axis describes how many sets of simulated observing runs have been completed in relation to the lifetime of the survey. The hashed contours correspond to the 68% spread of each of these statistics as measured by the 7 machine learning targeting algorithms.

of the lines at both sides of the figure is an artefact of the smoothing criteria chosen.

### 3.2 Target selection algorithms

In §3.1.1 we estimate the offset  $b_{z_{ML}}^{est}$  and the redshift error  $\delta_{z_{ML}}^{est}$  for the remaining potential targets. We now construct different selection algorithms to determine the optimal selection order for the subsequent simulated observing runs. We explore the following set of algorithms to determine which galaxies will be subsequently targeted, and summarise these selection algorithms in Table 1.

1) We choose N target galaxies in the date order that they were observed by the SDSS.

This represents one of the comparison methods. Again we note that the SDSS targeting algorithm which we are replicating in this method was not optimised for machine learning redshift estimation.

2) We choose target galaxies by randomly selecting N times without replacement from all remaining potential targets.

This strategy represents the second comparison method. It addresses the question; how well would we have performed, if we had selected targets at random from the list of potential candidates?

3) We choose target galaxies by selecting the N galaxies with the largest estimated value of redshift offset  $b_{z_{ML}}^{est}$ . We could expect this strategy to reduce the outlier fraction of a test sample of galaxies, because these cases will have high values of redshift offset.

4) We choose target galaxies by selecting the N galaxies with the largest estimated value of redshift error  $\delta_{z_{ML}}^{est}$ . This targeting strategy selects targets with the largest estimated redshift error. We could expect this strategy to reduce the value of the dispersion  $\sigma_{68}$  or  $\sigma_{95}$  of a test sample of galaxies.

5) We select target galaxies by selecting the  $N$  galaxies with the largest estimated value of the t-statistic  $T_s$ , which is defined as the ratio of offset to error,  $T_s = b_{z_{ML}}^{est} / \delta_{z_{ML}}^{est}$ . This targeting strategy examines the cases for which the estimated redshift error is in the least agreement with the estimated redshift offset. This algorithm can be viewed as selecting the extreme tails of the t-distribution.

6) We target galaxies by selecting the  $N$  galaxies with the largest estimated value of the harmonic mean  $H_m$  between offset and error, which is defined as,  $H_m = (b_{z_{ML}}^{est} \times \delta_{z_{ML}}^{est}) / (b_{z_{ML}}^{est} + \delta_{z_{ML}}^{est})$ . This targeting strategy gives equal weight to selecting galaxies with large values of both estimated redshift offset and estimated redshift error.

7) We select target galaxies by randomly drawing  $N$  galaxies from the binned distribution of the t-statistic  $T_s$ , with the probability of being selected proportional to the inverse of the number of data in the  $T_s$  bin. This targeting strategy begins by binning targets in the t-statistic values, and selects targets from the full distribution. However the random selection is chosen to give more weight to the outlying bins because these bins have the fewest number of objects. This statistic is chosen to determine if one can reduce both the outlier fraction and the  $\sigma_{68}, \sigma_{95}$  dispersions simultaneously.

8) We choose target galaxies by randomly drawing  $N$  galaxies from the binned distribution of the harmonic mean  $H_m$  between offset and error, with the probability of being selected again proportional to the inverse of the number of data in the  $H_m$  bin.

This targeting strategy again begins by binning targets in the values of the harmonic mean, and selects targets from the full distribution. The random selection is again chosen to give more weight to the outlying bins because these bins also have the fewest number of objects. This statistic is also chosen to determine if one can reduce both the outlier fraction and the  $\sigma_{68}, \sigma_{95}$  dispersions simultaneously.

9) We finally also generate a hybrid method which combines equal measures of randomly selected targets and targets with large estimated redshift errors. This final technique is motivated by our a posteriori observation that these techniques individually improve different sets of metrics, and therefore their combination may improve all metrics. In effect, the addition of random data acts to regularize the algorithm with large estimated redshift errors, so that it does not concentrate on only the worse performing target examples.

We note that we have also explored the use of a hybrid random selection method combined with targets which have large estimated redshift offsets, but did not find a noticeable difference between this method and method 9).

We measure how well each targeting algorithm compares to the two base strategies by estimating the machine learning photometric redshift for a distinct sample of test galaxies (see §2.1 for details). The training data for this machine learning redshift estimation is drawn from all galaxies up to and including the previous targeting run in each case.

### 3.3 Simulated observing runs

We examine the effect of the target algorithms on the recovered redshifts of an independent test sample by constructing an experiment to answer these questions:

1) What is the optimal ordering that the target galaxies should have been observed in, if we had wanted to improve the redshift estimates on a representative hold out sample.

2) How does this compare with the SDSS algorithm? We note that the SDSS algorithms were not designed with the above goal in mind.

3) How do the target algorithms compare with a random selection algorithm?

We ensure that each experiment within SDSS I&II or SDSS III have the same fixed pool of 100k target galaxies from which 1.3k galaxies are selected for each of the 75 simulated observing runs. Each algorithm will select targets from this fixed pool for follow up at different times. During each observing run we obtain a spectroscopic redshift for all of the selected target galaxies with 100% success. This ignores important observation effects such as fiber collisions, seeing conditions, and airmass. We commence each experiment after the first observing run has already been completed to provide an initial knowledge base to produce the initial machine learning predictions. We iteratively use all of the data acquired in previous runs to help decide which targets to observe in each subsequent run.

We will end the experiment when all of the galaxies have been targeted in the simulated sets of observing runs, irrespective of the order in which they were targeted. Therefore at the end of the experiment all the targets will have been selected by each of the different targeting algorithms, and we expect all of the final machine learning redshift results to converge.

We have also explored other follow up survey configurations, such as increasing or decreasing the number of observing runs (including 50, 100, 200), or the number of spectra taken during each run. We find similar improvements in final results in these cases once similar amounts of spectra have been collected.

We finally compare the algorithms to determine which methods produce the quickest reduction in error when using the targeted galaxies as training galaxies in a standard machine learning redshift estimation process. This is performed by passing the independent test data through the trained machine systems to first predict redshifts, and then by measuring performance metrics.

### 3.4 Estimating redshift performance on the test galaxies

In the previous sections we determined which targets should be selected during the next hypothetical observing run. We measure how well these different sets of selected galaxies are able to predict machine learning redshifts of a final test sample. For this purpose we use two extremely large test samples; one for the SDSS I&II analysis of size 769,479, and one for the SDSS III analysis of size 1,011,918, both of which are described in §2.1. These large test samples enable very accurate estimates of the metrics as measured on the point prediction redshift residuals  $\Delta_{z'} = (z_{spec} - z_{ML}) / (1 + z_{spec})$ , of the following derived statistical quantities; the median value

Acronym	Description
SDSS	Select targets in the date order by which they were observed by the SDSS.
Rand	Select targets at random.
LargeO	Select targets by the largest estimated redshift offset $b_{z_{ML}}^{est}$ .
LargeE	Select targets by the largest estimated redshift error $\delta_{z_{ML}}^{est}$ .
TsOE	Select targets by the largest estimated t-statistic: $(b_{z_{ML}}^{est}/\delta_{z_{ML}}^{est})$
HmOE	Select targets with largest estimated harmonic mean: $(b_{z_{ML}}^{est} \times \delta_{z_{ML}}^{est})/(b_{z_{ML}}^{est} + \delta_{z_{ML}}^{est})$
TsOESamp	Select targets by drawing from the binned t-statistic distribution with a probability of being selected to be inversely proportional to the number of data in the bin.
HmOESamp	Select targets by sampling from the binned estimated harmonic mean distribution as for TsOESamp.
LE_Rand	Select targets using a mixture of 50% random (Rand) and 50% of the largest estimated redshift error (LargeE).

**Table 1.** The different target selection routines explored in this work. The top two routines do not rely on machine learning, the other 7 routines use machine learning methods to estimate the redshift offsets  $b_{z_{ML}}^{est}$ , and the redshift errors  $\delta_{z_{ML}}^{est}$ , of the remaining targets. These values are then used in the specified way to select which targets should be subsequently observed in the next observing run.

$\mu$  of the distribution of  $\Delta_{z'}$ , the 68% and 95% spread of the distribution of  $\Delta_{z'}$  defined as  $\sigma_{68}(z')$ ,  $\sigma_{95}(z')$  and the outlier fraction defined by the percentage of galaxies for which  $|\Delta_{z'}| > 0.15$ .

When presenting the results of the different targeting algorithms we measure and present the ratio of relative improvement in each of the metrics, with respect to the time ordered SDSS targeting selection criteria. A negative improvement implies that the targeted galaxies produce worse metrics than those obtained by the SDSS targeting order.

This choice of metric is unstable for values which are relatively small and oscillate around zero. This is the case for the median of the residual distribution, which is less than 0.004 at 95%. We therefore do not examine the median value in what follows, and note that tree based methods often produce redshift estimates with low bias (see e.g. Carliles et al. 2008; Hoyle et al. 2015).

## 4 RESULTS

We begin by presenting the colour-magnitude distributions of galaxies which are selected by the different targeting algorithms with respect to the SDSS targeting algorithm. We then show how each targeting algorithm performs against the SDSS targeting order when estimating machine learning photometric redshifts of an independent test sample. We present results both as a function of the instantaneous value, corresponding to estimating redshifts as the survey is progressing, and as a function of the final value determined using all of the target galaxies. We then explore the effect of allowing a much enlarged target pool, which could mimic a change in follow-up strategy during the lifetime of the survey.

### 4.1 The distribution of targeted galaxies

In this section we examine which galaxies are preferentially selected by the different targeting algorithms. In Fig 3 we show which galaxies are selected using the largest photometric error (LargeE) target selection algorithm, for an increas-

ing number of simulated observing runs. The panels show the  $g - r$  colour against  $r$  band magnitude distribution after 5/75, 15/75 and 35/75 simulated sets of observations. We show the galaxies as observed in their original date order using the SDSS target algorithm by the red filled circles, and the LargeE targeting algorithm by the brown filled stars. In this figure we use the SDSS I&II data sample.

Examining the left-hand panel of Fig. 3 we find that initially the two distributions of galaxies are very different, with the LargeE algorithm preferentially selecting fainter and bluer galaxies. The experiment has been constructed such that the final sample of targeted galaxies is the same. Therefore, as expected, we find that the different colour-magnitude distributions of galaxies become more similar as we progress through the survey.

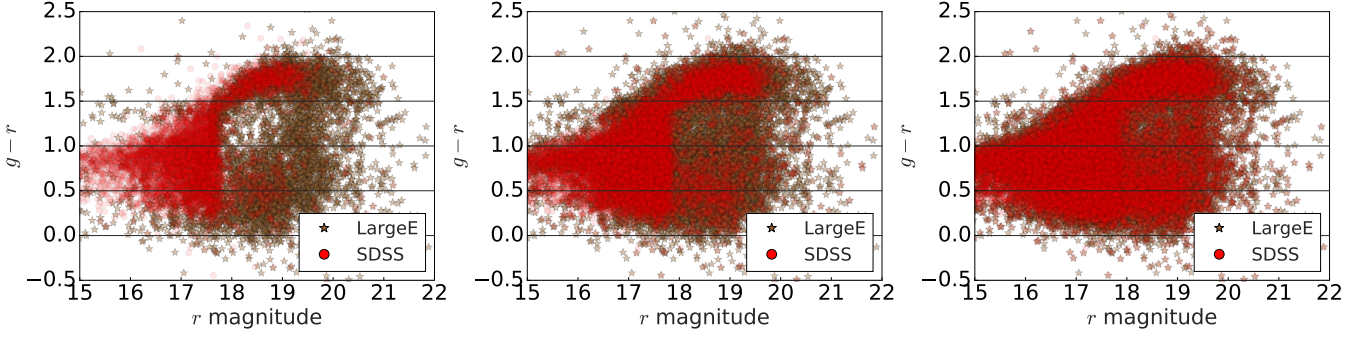
In Fig. 4 we show snapshots after five observing runs of the same galaxy distribution as Fig. 3 but for the following three different targeting algorithms from right to left: Random selection (Rand), Large Offset (LargeO), and the t-statistic (TsBE) between predicted bias and predicted error.

The panels of Fig. 4 shows that the random targeting algorithm and the t-statistic distribution targeting algorithm (TsBO) populate different regions of colour-magnitude space. The algorithm TsBO preferentially selects  $r$  band fainter galaxies across all colours, and a set of brighter, bluer galaxies. We note that the Large Bias algorithm closely resembles that of the Large Error targeting algorithm as seen in the left-hand panel of Fig. 3.

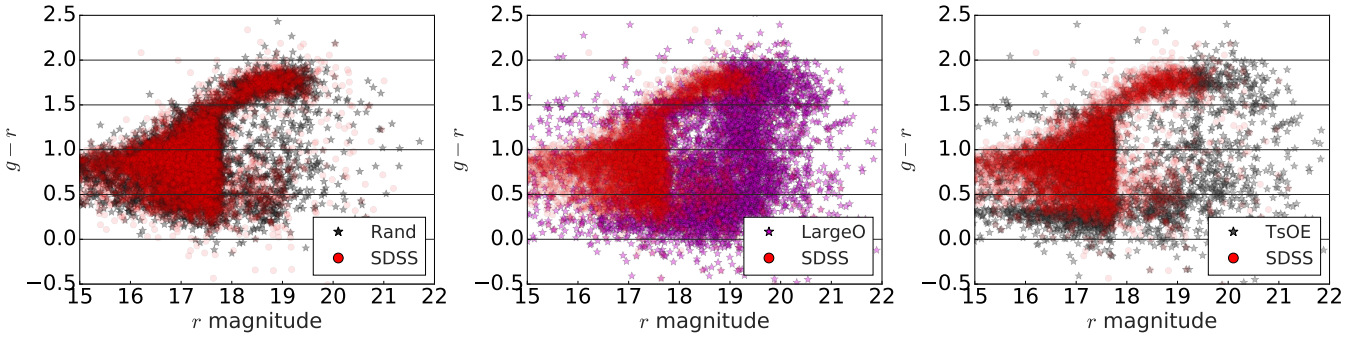
In Fig. 5 we choose to show the colour-magnitude distribution for the SDSS III sample, instead of the SDSS I&II sample as in the previous figures, of galaxies for the final three different targeting algorithms: the Harmonic mean of the predicted bias and predicted error (HmBE), the inverse number sampled t-statistic distribution TsBESamp, and the inverse number sampled Harmonic mean distribution (HeBESamp).

Examining Fig. 5 we find that both the sampled t-statistic and sampled Harmonic mean targeting algorithms produce samples of galaxies which are very similar to the inherent SDSS ordering. The Harmonic mean HmOE tar-





**Figure 3.** The  $g-r$  colour magnitude distribution of SDSS I&II galaxies selected by the SDSS and the Largest error (LargeE) target selection algorithms, after 5/75, 15/75, 35/75 (from left to right) simulated observing runs. Each observing run selects 1.3k galaxies.



**Figure 4.** The  $g-r$  colour magnitude distribution of SDSS I&II galaxies selected by the SDSS and the Random, Large predicted redshift bias, and large predicted t-statistic values. We show the different distribution of galaxies after 5/75 observing runs.

getting algorithm preferentially selects bluer galaxies at all  $r$  band magnitudes.

We note that in all of the figures in this section there are clearly different color-magnitude populations of galaxies being selected. These correspond to the different observing goals of both SDSS I&II and SDSS III, and correspond to main sample galaxies, luminous red galaxies, and quasars.

Finally in Fig. 6 we show the colour-magnitude distribution of galaxies using the hybrid random and large predicted redshift error (LargeE.Rand) algorithm for both the SDSS I&II (left-hand panel) and SDSS III (right-hand panel) after five simulated observing runs. In Fig. 6 we see that the colour-magnitude distribution of galaxies resembles that of both the random target selection (see the left-hand panel of Fig. 4), and the targeting algorithm using the predicted redshift error (see the left-hand panel of Fig. 3).

#### 4.2 The performance of different targeting algorithms

We show the effect of the different targeting algorithms using the metrics  $\sigma_{68}(z')$ ,  $\sigma_{95}(z')$ ,  $|\Delta_{z'}| > 0.15$  to describe photometric redshift performance as calculated on an independent test sample in Fig. 7 and Fig 8. We show SDSS I&II analysis in the left-hand panels and the SDSS III analysis in the right-hand panels. We show the relative improvement in each of the metrics, with respect to the time ordered SDSS targeting selection algorithm. We measure the relative improvement in each of these metrics with respect to two different SDSS values. In Fig. 7 we show the relative improve-

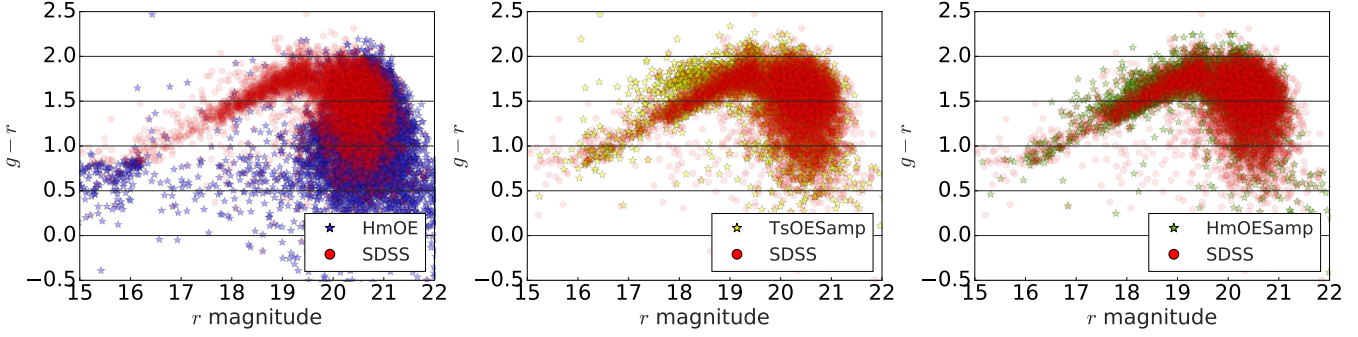
ment with respect to the instantaneous SDSS metric value, as calculated using the sample of galaxies which have been targeted up to and including that number of targeting runs. In Fig 8 we show the relative improvement with respect to the final value of the SDSS metric, which is calculated using the full sample of targeted galaxies, and is approximately the same as the final value from all other algorithms. In both figures the shaded contours around the random targeting algorithm (Rand) shows the 68% spread of that metric over 7 different realisations, and gives an estimate of the expected error on the other targeting routines. In each figure we always show the benchmark (SDSS), and (Rand) algorithms, and the overall best (LE.Rand) and worst (HmBESamp) performing algorithms, and then a random selection of two other algorithms. This choice is made for aesthetic considerations. We discuss each figure below.

##### 4.2.1 Comparison with the instantaneous SDSS metric values

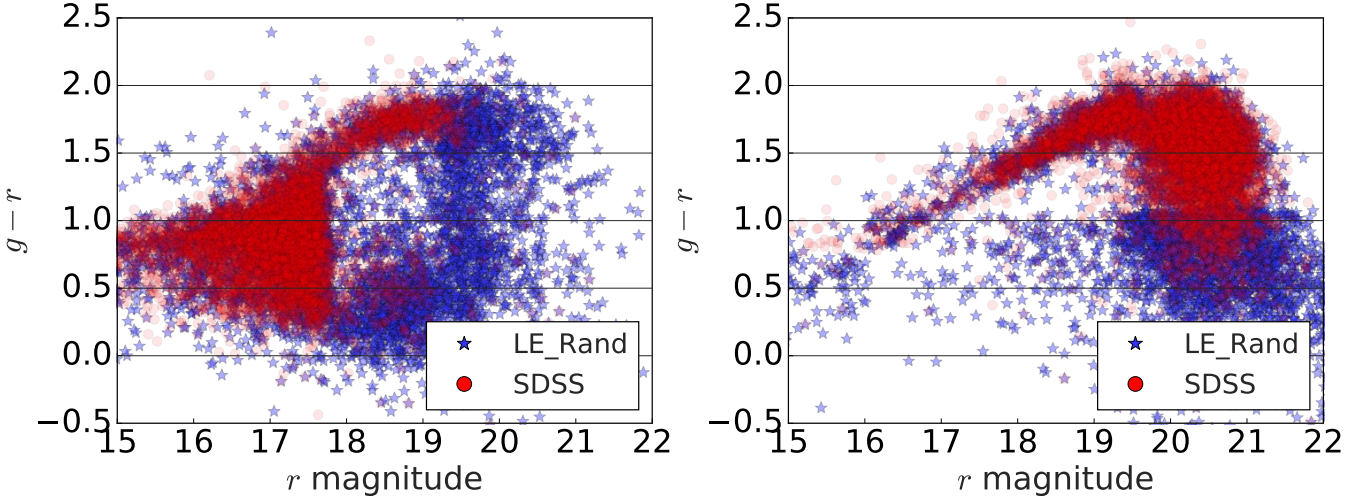
In the top panels of Fig. 7 we see that all of the targeting algorithms except for TsBESamp and LE.Rand perform less well than the random selection algorithm on the statistic  $\sigma_{68}(z')$  as measured on the independent redshift scaled residuals of the final test sample. This suggests how these results may apply to representative and unseen data. We note that the targeting algorithms not shown behave similarly to those presented.

The middle and bottom panels of Fig. 7 show how the targeting algorithms perform on the metrics  $\sigma_{95}(z')$  and





**Figure 5.** The  $g-r$  colour magnitude distribution of SDSS III galaxies selected by the SDSS and the HmBE, TsBESamp, and HmBESamp algorithms (from left to right) after 5/75 simulated observing runs.



**Figure 6.** The  $g-r$  colour-magnitude distribution of SDSS I&II (left-hand panel) and SDSS III (right-hand panel) galaxies using the SDSS and the hybrid LE\_Rand target selection algorithms. We again show the distribution of galaxies after 5/75 simulated observing runs.

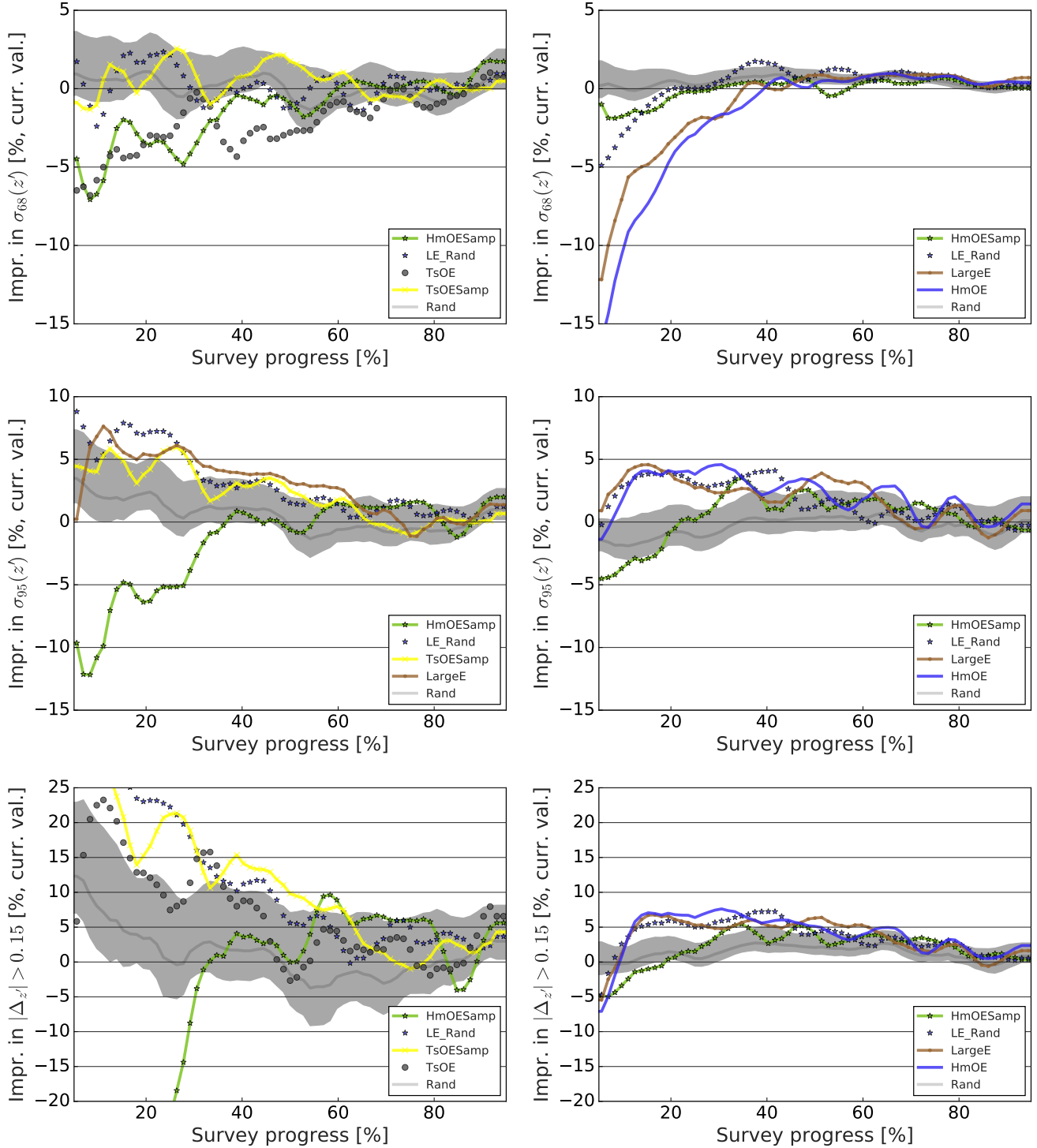
$|\Delta_{z'}| > 0.15$  as calculated on the independent test sets. We find that almost all of the target selection algorithms either slightly, or moderately, out perform the random selection algorithm after approximately 20% of the simulated observing runs have been completed. This improvement can be expected because many of the targeting algorithms, including LargeE, select galaxies which are predicted to have large estimated values of redshift offset and redshift error, see also §3.2. By preferentially selecting these galaxies we expect the wings of the distributions of  $\Delta_{z'}$  to be removed. This also explains why the improvement with respect to the random algorithm increases as we pass from  $\sigma_{95}(z')$  to  $|\Delta_{z'}| > 0.15$ , which effectively samples a larger component of the wings of the distribution.

Finally we note that as the survey surpasses the 60% completion mark, all of the algorithms converge. We can understand this by recalling that this experiment has been performed by imposing a fixed list of potential targets, and each algorithm will eventually have selected the same target galaxies. We explore this effect further in §4.2.3 by relaxing this constraint. We note that in all panels in Fig. 7 we find that the targeting algorithm HmBESamp performs worse

than all the other sampling algorithms including that of the random sampling (Rand).

Examining all panels of Fig. 7 we see that the random targeting algorithm is always close to 0% improvement compared to the SDSS, which implies consistency with the measured value of each metric between both algorithms. This suggests that the SDSS target algorithm selected galaxies randomly once target lists were made available to them. This conclusion is consistent with their targeting strategy (see Strauss et al. 2002).

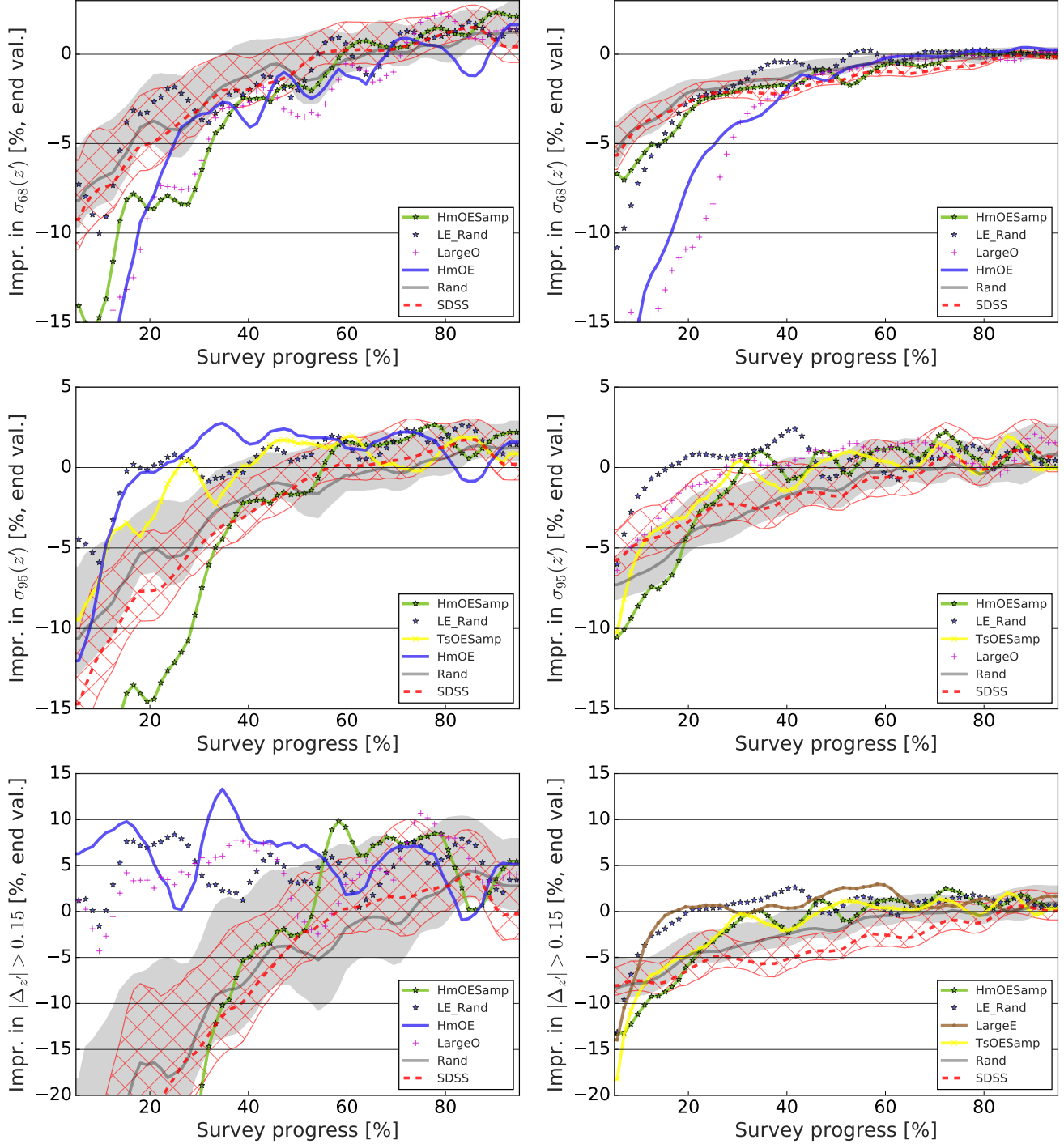
If we concentrate on the LE\_Rand algorithm in each panel of Fig. 7, we see that for a given number of target runs, this targeting algorithm either does no worse than, or improves by 5-25%, the measured metrics when compared with the SDSS and Rand targeting algorithms. This motivates the a priori choice for the construction of this targeting algorithm made in §3.2. This enhancement is less pronounced for the SDSS III sample, but there is still improvement with respect to both the SDSS and the random targeting selection algorithms. The distribution of SDSS III galaxies is inherently redder and more homogeneous than that of the SDSS I&II, e.g., see Fig. 6. This explains why the improvement in the metrics is less pronounced as more data is added to



**Figure 7.** The relative improvement of each measured metric (see y-axis label) value for different targeting algorithms with respect to the instantaneous (or current, curr.) value from the SDSS time ordered targeting algorithm. The left-hand (right-hand) panels show the results for the SDSS I&II (SDSS III) analyses. The shaded region corresponds to the 68% spread of metric values using the random targeting algorithm (Rand) across 7 independent experiments.

the machine learning systems. We conclude that if the SDSS had wanted to optimise their target selection algorithm, to improve a machine learning redshift estimation of a final test sample, *and* the full list of potential targets were available apriori, they would have been able to improve their

redshift metrics by 5-25% using the suggested LE\_Rand selection algorithm. These improved redshift estimates could have potentially lead to earlier science results for some science applications.



**Figure 8.** The relative improvement of each measured metric (see y-axis label) value for different targeting algorithms with respect to the final (or end) value from the SDSS date ordered targeting algorithm. The left-hand (right-hand) panels show the results for the SDSS I&II (SDSS III) analyses. The dark grey shaded region corresponds to the 68% spread of metric values using the random targeting algorithm (Rand) across 7 independent experiments, and the red dashed lines and hatched regions corresponds to the spread of metric values using the SDSS algorithm.

#### 4.2.2 Comparison with the final SDSS metric values

We next present the same analysis as in the previous subsection, but show the improvement in the measured metrics with respect to the final value obtained at the end of all of the observing runs. We expect all of the routines to converge

at the end of the survey, because they have exhaustively selected the target list. This method of presentation allows us to also explore the improvement of metrics using the SDSS targeting algorithm as more training galaxies are added, as shown in Fig. 8, however we note that this result has been present before (e.g. Hoyle et al. 2015), within the context of

randomly selecting training data. The red dashed lines and hatched region correspond to the median and 68% dispersion for each metric as measured using the galaxies targeted using the SDSS algorithm. The left-hand panels again show the results of the SDSS I&II analysis and the right-hand panels show the results of the SDSS III analysis.

Examining the top panels of Fig. 8 we find that all of the algorithms require at least 50% of the available data before the measured values of  $\sigma_{68}(z')$  approach the final value as measured by the SDSS. We note that the values at the right edge of many of the panels are slightly above the 0% improvement mark, which is due to the smoothing and dispersion of the results which causes the final value of the metrics to be biased slightly low. We find that the hybrid targeting algorithm LE\_Rand again performs well compared to both the random (Rand) and the time ordered SDSS targeting algorithms, and some of the other targeting algorithms, for example, LargeE, TsBE and LargeB all perform substantially worse than the random algorithm until around 50% of the runs have been completed in most metrics. Examining the bottom two rows of Fig. 8 we find that the performance of many of the targeting algorithms, including LE\_Rand, LargeE, and LargeO show an improvement over both the random algorithm, and the SDSS algorithm. In particular after about 20% of the survey has been completed, these algorithms already produce values which are consistent with the final estimates for both of the metrics  $\sigma_{95}(z')$  and the outlier fraction  $|\Delta_{z'}| > 0.15$ . This can be compared to the amount of the survey 60% (80%) which is required by the SDSS and random algorithms for the SDSS I&II (III) targeted galaxies before obtaining metric values which are consistent with the final values.

The values of the measured statistics approach the final values after approximately 40-60% of the simulated survey runs, depending on the target selection algorithm. This suggests that the information content of the training set is already saturated for this fixed target pool. In the next subsection we allow a much enlarged target pool to be drawn from. The motivation for this is that one could potentially change survey strategies once the redshift estimates stabilise.

#### 4.2.3 Enlarged potential candidate list

Rather than being constrained to the same samples of target candidates as in §4.2.1 & §4.2.2, we now allow the machine learning targeting algorithms to select targets from a much larger target pool. This could be realised in practise by noting that some of the algorithms, such as LE\_Rand produce good redshift estimates within the first 40% or 50% of the available survey time. One could then change follow-up survey strategies and construct a new target list drawn from fainter galaxies, which require longer observing times or resources, or new target lists from an enlarged survey footprint.

To explore this approach we generate two training datasets labelled ‘tr1’ and ‘tr2’, and one test data sample in each of the separate SDSS I&II and SDSS III analysis. The sizes of tr1 and the test sample are fixed to 100k. The size and selection of tr2 corresponds to all galaxies which are not in the test sample and is of size 770k for the SDSS I&II analysis and 1011k for the SDSS III analysis. We note

that this implies that tr1 is a subsample of tr2, and both tr1 and tr2 are independent from the test samples.

We restrict the random and SDSS targeting algorithms to the smaller data set tr1, but allow the other targeting algorithms the freedom to select from the larger pool of potential target galaxies tr2. In Fig. 9 we present the relative improvement of each metric (see the y-axis labels) value for different targeting algorithms with respect to the final value from the SDSS time ordered targeting algorithm.

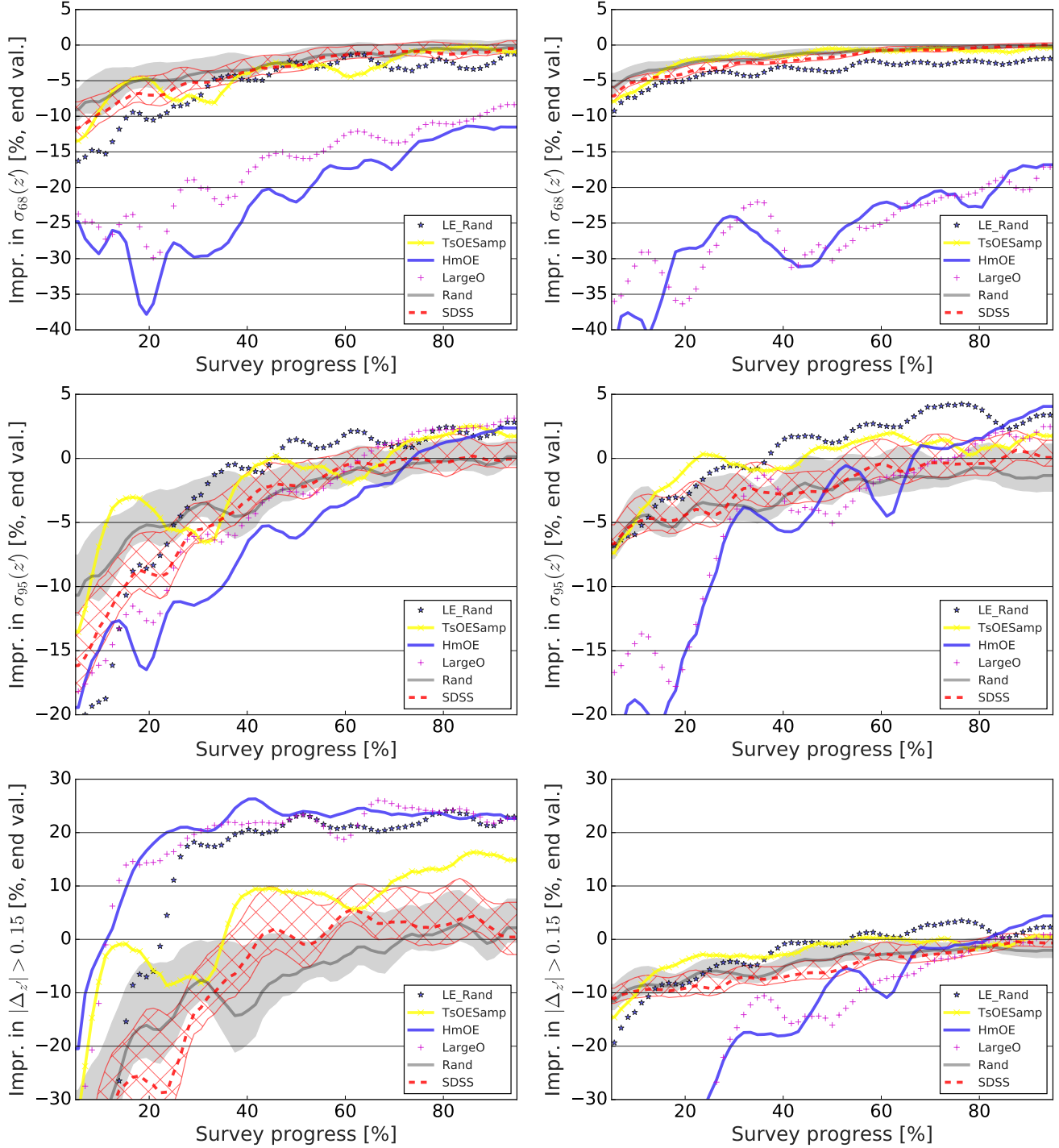
By examining all panels of Fig. 9 we again find that after approximately 20% of the observing runs have been taken, the hybrid target selection algorithm (LE\_Rand) performs as least as well and often better on all metrics in both sets of analysis than both the random algorithm, and the SDSS algorithm. We remind the reader that the SDSS algorithm was not designed to be optimised for machine learning redshift estimation, and therefore a direct comparison is included here for completeness. We find that the two other target selection algorithms LargeB and HmBE perform poorly on the metrics  $\sigma_{68}(z')$ , less poorly on the metric  $\sigma_{95}(z')$ , and perform very well on the outlier fraction metric  $|\Delta_{z'}| > 0.15$ . This suggests that these algorithms are targeting galaxies which remove the outliers from the tails of the distribution, but are not improving the redshift estimates of the majority of the test sample galaxies. We note that the improvement of the hybrid algorithm LE\_Rand is much less pronounced for the SDSS III sample than for the SDSS I&II sample, which is probably due to the more homogeneous sample of SDSS III galaxies.

One can directly see the effect of allowing an enlarged target pool by examining both Figs 8 & 9. The random and SDSS lines and contours are equivalent in both figures, and the improvement in the metric values of some of the machine learning target selection algorithms is noticeable.

## 5 SUMMARY & CONCLUSIONS

Photometric galaxy catalogues can be maximally exploited for cosmological analyses once galaxy redshifts have been measured spectroscopically, or estimated photometrically. Machine learning architectures can map measured photometric features onto spectroscopic redshifts for a subset of ‘training’ data with both quantities. This mapping can then be used to estimate redshifts for all photometrically selected galaxies which are representative of the training sample. Constructing a training set often requires programs of dedicated telescope time to perform spectroscopic follow up of a set of target galaxies for which one may reasonably assume a redshift can be measured. In this paper we propose a machine learning based target selection algorithm which leads to faster improvements in the machine learning redshift estimation of a final sample of test galaxies, such that less spectra are required.

We explore 7 different target selection algorithms, in addition to the SDSS, and a random target selection algorithm which are constructed by applying machine learning techniques in a subtly different way than normal redshift estimation. Instead of estimating the value of the photometric redshift we construct a machine to predict the size of the photometric redshift error and the photometric offset, or bias, defined as  $|z_{spec} - z_{ML}|$ . We find that these



**Figure 9.** The relative improvement of each measured metric (see the y-axis labels) value for different targeting algorithms with respect to the final (or end) value from the SDSS date ordered targeting algorithm. The left-hand (right-hand) panels show the results for the SDSS I&II (SDSS III) analyses. The grey shaded region corresponds to the 68% spread of metric values using the random targeting algorithm (Rand) across 7 independent experiments, and the red dashed lines and hatched regions corresponds to the spread of metric values using the date ordered SDSS targeting algorithm. In this figure only, the machine learning targeting algorithms may draw targets from a much larger target pool than the SDSS and Rand algorithms.

predictions are very accurate, for example  $(2 \pm 5) \times 10^{-3}$  for photometric redshift errors, which, for SDSS data, typically have values greater than  $2 \times 10^{-2}$ . We combine these predictions to construct 7 different machine learning target selection algorithms. We find that a hybrid algorithm which

selects equal numbers of random galaxies (denoted by ‘Rand’ in Table. 1), and galaxies with the largest predicted redshift error (LargeE) provides the best target selection routine of those examined.

To test our methods we construct sets of experiments



using data drawn from the Sloan Digital Sky Survey Data Release 12 (SDSS DR12 Alam et al. 2015). We perform two independent sets of analyses by selecting data observed within the time frames of SDSS I&II and SDSS III. For the purposes of this work, we construct clean samples of galaxies, free of stellar contamination. One may also choose to select combinations of stars, galaxies, and other artefacts if the science goal were to improve star/galaxy separation using machine learning.

To compare the different target selection methods we compute photometric redshifts on an independent sample of representative galaxies and measure the following metrics:  $\sigma_{68}(z')$  which is defined to be the 68% dispersion of the redshift scaled residuals  $\Delta_{z'} = (z_{spec} - z_{ML})/(1 + z_{spec})$ , and would be equivalent to 1 standard deviation if  $\Delta_{z'}$  were well described by a Gaussian; and the values of the 95% dispersion of  $\Delta_{z'}$  defined as  $\sigma_{95}(z')$ ; as well as the outlier fraction defined by the fraction of data for which  $|\Delta_{z'}| > 0.15$ .

We compare the 7 targeting algorithms presented in this work with a random target selection algorithm, and with the SDSS targeting order as defined by the date when the spectra were obtained. We find that after 20% of simulated runs, the hybrid targeting algorithm performs no worse than the random and SDSS targeting algorithms on the metric  $\sigma_{68}(z')$ . We find that after only 10% of the total number of observing runs the value of  $\sigma_{95}(z')$ , and the outlier fraction defined are between 5% and 20% better than both a random selection algorithm, and that of the SDSS. We note that the SDSS target selection algorithms (see York & SDSS Collaboration 2000; Eisenstein et al. 2001; Richards et al. 2002) were not constructed for optimal machine learning redshift estimation, and therefore a direct comparison between the SDSS and the targeting algorithms described in this work is for presentation purposes only. It however shows that machine learning based target selection algorithms can be used to tailor spectroscopic follow-up for specific science goals.

We also estimate how many observing runs are required before the metric values, which are measured on the independent test samples, approach the final metric values, as calculated using all of the target galaxies. We find that the hybrid method is able to reach the same precision in the outlier fraction and  $\sigma_{95}(z')$  within 20% of the total amount of followed-up targets. To obtain similar precision on the value of  $\sigma_{68}(z')$  one requires between 40% (60%) of observed targets for the SDSS I&II (III) data sample.

Assuming that one may reach the same precision with less observing runs, suggests that this method could be used to change observing strategies for the latter part of a follow up survey, in order to target different samples of galaxies. This could be used to increase the footprint of the survey, which could help with, for example, clustering redshift estimates (see Menard et al. 2013), or to probe to deeper apparent magnitude limits which require longer integration times. Of course such a target selection algorithm would also have to be convolved with the science goals of the survey. For example for Baryon Acoustic Oscillation surveys, one requires large area samples of homogeneous galaxies.

We estimate how a difference in the selected targets could effect the recovered metrics, by allowing all of the machine learning targeting routines, i.e. except the time ordered SDSS and random algorithms, to draw targets from

a much enlarged pool of target galaxies. Using the hybrid targeting algorithm we find that the outlier fraction is improved by 20% compared to that of the SDSS or random algorithms. A further extension of this work would be to change the test sample, as the target sample becomes enlarged, allowing the test sample to cover a larger range of photometric features values. This will be explored in future work.

During this work we have made a few simplifying assumptions. For example we have assumed that each potential target could be observed in the run of our choice. In practice this may not be possible due to observing conditions such as the galactic coordinates of each target, or due to fiber collisions for targets which are nearby on the sky. Instead one could prepare a list of potential targets which could be viewed in the subsequent target run, and then determine which of these targets would be the best to follow up. Furthermore most surveys are not designed to have a purely photometric phase which would identify all target candidates, followed by a dedicated spectroscopic follow up. In these cases the techniques presented in this paper could still be applied, by introducing and varying an ever increasing set of potential targets as more photographic data is obtained. Should one wish to use this technique for surveys that need to define a spectroscopic target selection before initial imaging or spectra have been taken, one may augment the dataset using existing imaging of comparable depth, or using realistic simulations, and perform a similar analysis described in this work.

One import point to note is that in this work we have assumed that the sample of galaxies with spectra is representative of those galaxies without spectra, and that test sample spectra may be obtained. This is not necessarily the case in reality due to, for example, the difficulty of obtaining spectra for very faint galaxies or objects which sit in the redshift desert. A future approach to explore this problem could be to bias the initial training sample by applying sets of color cuts and repeating this analysis. Then we would have to further impose that the selected sample for spectroscopic follow-up remains biased in some sense (perhaps in brightness or color) with respect to a final test sample. The exact nature of the biases we should incur would vary for each observation survey.

Other extensions could be to combine the estimation of both the bias,  $b_{z_{ML}}$  and error  $\delta_{z_{ML}}$  simultaneously, not separately, as performed here (see §3.1.1). Many machine learning algorithms, including decision tree based methods allow such a multiple output feature training. Furthermore we have not explored the use of other machine learning algorithms to perform the prediction or redshift estimation tasks. One could extend this work by exploring different algorithms. We choose to use random forests because they are easily run in parallel across many cores, and they are inherently very fast because they use decision trees. Random forests are also very robust machine learning frameworks which perform well compared with other methods on many problems in the literature (e.g., Hildebrandt et al. 2010; Sánchez et al. 2014). One could also extend this analysis by comparing the full redshift distribution of the photometric and spectroscopic samples, instead of the metrics chosen in this paper.

The target selection methods presented in this paper

could be applied to the Dark Energy Survey (The Dark Energy Survey Collaboration 2005) or Euclid (Laureijs et al. 2011) by first defining a training sample and using this to predict which target galaxies will have large redshift errors and biases in a machine learning approach, and then iteratively following up sub samples of galaxies and retrain the prediction models. It could also be applied with other science goals in mind, for example to star and galaxy separation, or to identify samples of galaxies within a particular population class, or redshift range (for example see <sup>4</sup>).

## A CASJOBS MYSQL QUERY

We obtain observational data from the SDSS using the following MySQL query which is run in the DR12 schema:

```
select s.specObjID, q.objid,
q.dered_u, q.dered_g, q.dered_r, q.dered_i, q.dered_z,
q.modelMagErr_u, q.modelMagErr_g, q.modelMagErr_r,
q.modelMagErr_i, q.modelMagErr_z,
q.petroRad_r,
q.type as photpType,
s.z as specz, s.zerr as specz_err,
s.zWarning, s.mjd
into mydb.specPhotoDR12v3 from SpecObjAll as s
join photoObjAll as q on s.bestobjid=q.objid
```

This produces 3.9 million results of which we further process as described in §2.1.

## ACKNOWLEDGMENTS

The authors would like to thank an anonymous referee for comments and suggestions which have improve the paper. B. Hoyle, S.Seitz and M. M. Rau acknowledge support from the Transregional Collaborative Research Centre TRR 33 - The Dark Universe and the DFG cluster of excellence “Origin and Structure of the Universe”. Funding for the SDSS and SDSS-II has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, the U.S. Department of Energy, the National Aeronautics and Space Administration, the Japanese Monbukagakusho, the Max Planck Society, and the Higher Education Funding Council for England. The SDSS Web Site is <http://www.sdss.org/>.

## REFERENCES

- Alam S., Albareti F. D., et al. 2015, ArXiv e-prints:1501.00963  
 Bonnett C., 2015, MNRAS, 449, 1043  
 Breiman L., 2001, Machine Learning, 45, 5  
 Breiman L., Friedman J. H., Olshen R. A., Stone C. J., 1984, Classification and Regression Trees. Wadsworth International Group, Belmont, CA

- Carliles S., Budavári T., Heinis S., Priebe C., Szalay A., 2008, in Argyle R. W., Bunclark P. S., Lewis J. R., eds, Astronomical Data Analysis Software and Systems XVII Vol. 394 of Astronomical Society of the Pacific Conference Series, Photometric Redshift Estimation on SDSS Data Using Random Forests. p. 521  
 Carrasco Kind M., Brunner R. J., 2013, MNRAS, 432, 1483  
 Collister A. A., Lahav O., 2004, PASP, 116, 345  
 Csabai L., Dobos L., Trencsényi M., Herczegh G., Józsa P., Purger N., Budavári T., Szalay A. S., 2007, Astronomische Nachrichten, 328, 852  
 Dieleman S., Willett K. W., Dambre J., 2015, ArXiv: 1503.07077  
 Eisenstein D. J., Annis J., Gunn J. E., Szalay A. S., Connolly A. J., Nichol R. C., et al., 2001, AJ, 122, 2267  
 Eisenstein D. J. e. a., 2011, AJ, 142, 72  
 Gerdes D. W., Sypniewski A. J., McKay T. A., Hao J., Weis M. R., Wechsler R. H., Busha M. T., 2010, ApJ, 715, 823  
 Gunn J. E., Siegmund W. A., Mannery E. J., Owen R. E., Hull C. L., Leger R. F., Carey L. N., Knapp G. R., York D. G., Boroski W. N., Kent S. M., Lupton R. H., Rockosi C. M., et al., 2006, AJ, 131, 2332  
 Hála P., 2014, ArXiv e-prints:1412.8341  
 Hastie T., Tibshirani R., Friedman J., 2001, The Elements of Statistical Learning. Springer Series in Statistics, Springer New York Inc., New York, NY, USA  
 Hildebrandt H., Arnouts S., Capak P., Moustakas L. A., Wolf C., Abdalla e. a., 2010, Astron. & Astrophys., 523, A31  
 Hogan R., Fairbairn M., Seeburn N., 2015, MNRAS, 449, 2040  
 Hoyle B., 2015, ArXiv:1504.07255  
 Hoyle B., Rau M. M., Bonnett C., Seitz S., Weller J., 2015, MNRAS, 450, 305  
 Hoyle B., Rau M. M., Paech K., Bonnett C., Seitz S., Weller J., 2015, MNRAS, 452, 4183  
 Hoyle B., Rau M. M., Zitlau R., Seitz S., Weller J., 2015, MNRAS, 449, 1275  
 Jouvel S., Abdalla F. B., Kirk D., Lahav O., Lin H., Annis J., Kron R., Frieman J. A., 2014, MNRAS, 438, 2218  
 Kohonen T., ed. 1997, Self-organizing Maps. Springer-Verlag New York, Inc., Secaucus, NJ, USA  
 Lahav O., 1997, in Di Gesu V., Duff M. J. B., Heck A., Maccarone M. C., Scarsi L., Zimmerman H. U., eds, Data Analysis in Astronomy Artificial neural networks as a tool for galaxy classification.. pp 43–51  
 Laureijs R., Amiaux J., Arduini S., Auguères J. ., Brinchmann J., Cole et al. 2011, ArXiv e-prints:1110.3193  
 Li N., Thakar A. R., 2008, Computing in Science and Engineering, 10, 18  
 Masters D., Capak P., Stern D., Ilbert O., et al., 2015, ApJ, 813, 53  
 Menard B., Scranton R., Schmidt S., Morrison C., Jeong D., Budavari T., Rahman M., 2013, ArXiv: 1303.4722  
 Polsterer K. L., Gieseke F., Igel C., Goto T., 2014, in Manset N., Forshay P., eds, Astronomical Data Analysis Software and Systems XXIII Vol. 485 of Astronomical Society of the Pacific Conference Series, Improving the Performance of Photometric Regression Models via Massive Parallel Feature Selection. p. 425  
 Rau M. M., Seitz S., Brimiouille F., Frank E., Friedrich O.,

<sup>4</sup> [sagasurvey.org](http://sagasurvey.org)



- Gruen D., Hoyle B., 2015, MNRAS, 452, 3710
- Richards G. T., Fan X., Newberg H. J., Strauss M. A., Vanden Berk D. E., et al., 2002, AJ, 123, 2945
- Sánchez C., Carrasco Kind M., Lin H., Miquel R., Abdalla F. B., et al., 2014, MNRAS, 445, 1482
- Smith J. A., et al., 2002, AJ, 123, 2121
- Strauss M. A., Weinberg D. H., Lupton R. H., Narayanan V. K., Annis J., et al., 2002, AJ, 124, 1810
- Tagliaferri R., Longo G., Andreon S., Capozziello S., Donalek C., Giordano G., 2003, Lecture Notes in Computer Science, 2859, 226
- The Dark Energy Survey Collaboration 2005, ArXiv: 0510346
- Vanzella E., Cristiani S., Fontana A., Nonino M., Arnouts S., Giallongo E., Grazian A., Fasano G., Popesso P., Saracco P., Zaggia S., 2004, Astron. & Astrophys., 423, 761
- Yeche C., Petitjean P., Rich J., Aubourg E., Busca N., Hamilton J. ., Le Goff J. ., Paris I., Peirani S., Pichon C., Rollinde E., Vargas-Magana M., 2009, ArXiv: 0910.3770
- York D. G., SDSS Collaboration 2000, AJ, 120, 1579